

PENGENALAN KARAKTER MANDARIN DENGAN BACKPROPAGATION NEURAL NETWORK

Chairisni Lubis¹⁾, Tony²⁾, Ardy Kuncoro³⁾

^{1), 2), 3)} Fakultas Teknologi Informasi Universitas Tarumanagara
Jalan S.Parman No.1 Jakarta 11140

Email : chairisni.fti.untar@gmail.com ¹⁾, tony@fti.untar.ac.id ²⁾, liangzai992@gmail.com ³⁾

Abstrak

Mandarin merupakan bahasa kedua dunia setelah bahasa Inggris. Karakter Mandarin sukar dipelajari karena terbentuk dari beberapa goresan (stroke) dengan jumlah goresan yang bervariasi. Pada penelitian ini, Backpropagation Neural Network (BPNN) digunakan untuk mengenali tulisan tangan karakter mandarin (kanji). Ada beberapa pemrosesan awal yang digunakan untuk mendapatkan hasil pengenalan yang maksimum diantaranya : *smoothing*, *automatic cropping*, dan *feature extraction*. Tingkat keberhasilan pengenalan maksimum yang dicapai pada penelitian ini sebesar 59,2 %.

Kata kunci: Automatic Cropping, Backpropagation Neural Network, Global Histogram.

1. Pendahuluan

Mandarin adalah salah satu bahasa yang mulai menunjukkan keberadaan dan peranannya di dunia. Keseluruhan dari huruf mandarin diperkirakan lebih dari 70.000 kata [1]. Huruf-huruf mandarin tidak sama seperti alphabet, karakter mandarin adalah gabungan dari goresan goresan dasar yang digabungkan sehingga membentuk sebuah karakter mandarin Pada awalnya mandarin merupakan bahasa yang sukar untuk dipelajari, oleh karena itu dengan aplikasi ini diharapkan dapat membantu untuk memudahkan dan juga membantu pembaca untuk dapat mengerti huruf Mandarin.

Pada penelitian ini, karakter mandarin hasil tulisan tangan akan dicoba untuk dikenali dengan menggunakan *Backpropagation Neural Network* (BPNN). Ciri dari setiap karakter akan didapat melalui proses ekstraksi ciri *Global Histogram*, yang merupakan input bagi BPNN.

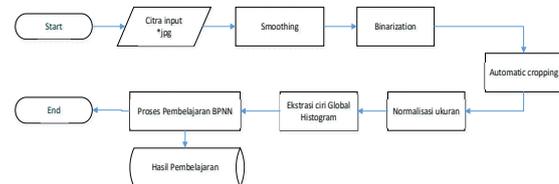
2. Pembahasan

Pada pengenalan karakter Mandarin dengan menggunakan BPNN ini, selain menggunakan kedua metode di atas, digunakan juga beberapa metode untuk perosesan awalnya (*preprocessing*) yaitu :

- *Gaussian Filter* : digunakan untuk menghaluskan citra
- Metode Otsu : digunakan untuk proses binerisasi
- *Automatic Cropping* : digunakan untuk proses cropping secara otomatis

- Normalisasi : digunakan untuk menyeragamkan dimensi citra.

Sebelum dilakukan proses pengenalan dilakukan dahulu proses pembelajaran dengan menggunakan BPNN. Diagram alir (*flowchart*) pembelajaran BPNN dapat dilihat pada gambar 1. Pada proses pengenalan digunakan juga diagram alir yang sama, tetapi dengan menggunakan bobot keterhubungan dari hasil pembelajaran. Input yang berupa tulisan tangan karakter Mandarin berasal dari hasil *scanning* berformat .jpg.



Gambar 1. Flowchart Proses Pembelajaran BPNN

2.1 Karakter Mandarin

Huruf atau karakter Mandarin merupakan satu di antara huruf tertua yang dikenal di dunia. Bahasa Mandarin (Tradisional: 北方話, Sederhana: 北方话, Hanyu Pinyin: Běifānghuà, artinya: “bahasa percakapan Utara” adalah dialek Bahasa Tionghoa yang dituturkan di sepanjang utara dan barat daya Republik Rakyat China. Kata “Mandarin”, dalam bahasa Inggris digunakan untuk menerjemahkan beberapa istilah China yang berbeda yang merujuk kepada kategori-kategori bahasa China lisan.

Huruf-huruf Mandarin tidak sama seperti alphabet, huruf Mandarin didasari atas ideograf dan piktograf. Tulisan Mandarin merupakan sebuah gambar yang dijumpai 8000 tahun yang lalu berdasarkan objek atau benda. Setelah mengalami evolusi, barulah gambar tulisan Mandarin menjadi sistem penulisan yang lengkap diciptakan sekitar 3000 tahun yang lalu di china, sehingga menjadikan salah satu cara penulisan tertua di dunia. [2]

Huruf Mandarin terbagi menjadi dua jenis yaitu komponen tunggal dan komponen gabungan. Sebagian besar huruf Mandarin adalah komponen gabungan di berbagai posisi huruf tersebut. Contohnya 木(mù) “pohon”, dan jika dua aksara 木 diletakkan bersama pada sisi masing-masing, maka menjadi 林 (lín)

“hutan” . Penggabungan 日 (ri) “matahari” dan 月 (yuè) “bulan” membentuk 明 (míng) “terang” . Bahasa Mandarin adalah bagian dari kelompok bahasa China yang dalam skala lebih luar merupakan bagian dari rumpun bahasa Sino-Tibet. Bahasa Mandarin mempunyai 4 nada, setiap kata yang mempunyai bunyi yang sama namun nadanya berbeda dapat berbeda pula artinya. [7]

Huruf Mandarin yang digunakan sekarang ini berasal dari pictograph (tulisan yang menggunakan lambang) yang diukir di atas tulang-tulang yang berisi kata-kata bijaksana berasal lebih dari 3.000 tahun yang lalu, dan pictograph yang ditemukan beberapa lama sebelumnya. Dalam perkembangan sejarahnya, huruf Mandarin berubah dari pictograph menjadi huruf yang tersusun dari goresan-goresan dengan struktur yang lebih jauh sederhana.

Terdapat juga jumlah garis (stroke) yang sering ditemui dan menjadikan bahasa Mandarin sukar untuk dipelajari, teknis menulis bahasa Mandarin tidak boleh asal-asalan, ada teknik dan aturannya. Terdapat 8(delapan) stroke dasar yang kemudian dimodifikasi menjadi berbagai variasi untuk membentuk sebuah karakter Mandarin yang serumit apapun. Untuk mengetahui suatu karakter memiliki berapa stroke adalah dengan melihat berapa goresan dasar yang digabungkan untuk membuat sebuah karakter Mandarin. [9]

Manusia memiliki cara penulisan yang ciri-ciri dalam penulisan yang berbeda. Oleh sebab itu, tulisan tangan menjadi kendala dalam pengolahan data, karena komputer tidak dapat mengenali citra tulisan tangan, maka dari itu sebuah sistem akan dikembangkan untuk mengelolah dan mengenali sebuah data citra tulisan tangan karakter Mandarin untuk dikenali oleh komputer.

2.2 Gaussian Filter

Penghalusan citra (*Smoothing*) bertujuan menurunkan atau menekan *noise* pada suatu citra. *Noise* pada citra tidak hanya terjadi karena ketidak sempurnaan dalam proses pengambilan, tetapi bias juga disebabkan oleh adanya kotoran pada citra. Proses *smoothing* ini dilakukan dengan cara mengganti setiap titik (x(t),y(t)) di lintasan dengan rata-rata nilai tetangganya [3].

Untuk mendapatkan citra hasil *smoothing* (G(x,y)) dapat digunakan persamaan berikut :

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \dots(1)$$

Keterangan :
 σ : nilai sebaran

dan untuk proses *masking* digunakan mask 5x5 seperti yang terdapat pada tabel 1 di bawah ini [10].

Tabel 1. *Mask Gaussian Filter 5x5 (σ=1,4)*

0,01	0,022	0,029	0,022	0,01
0,022	0,048	0,062	0,048	0,022
0,029	0,062	0,081	0,062	0,029
0,022	0,048	0,062	0,048	0,022
0,01	0,022	0,029	0,022	0,01

2.3 Binerization

Citra biner adalah citra yang hanya mempunyai dua nilai yaitu hitam dan putih. Proses binerisasi menghasilkan citra biner. Piksel-piksel objek menjadi warna hitam yang memiliki nilai 1, dan latar belakang menjadi warna putih yang memil nilai 0. Citra biner dapat dilihat pada Gambar 2.



Gambar 2. *Citra Biner*
 Sumber : Logo, Citra Biner
<http://elektronika-dasar.web.id/teori-elektronika/definisi-dan-pengolahan-citra-digital/>, 10 September 2014

Permasalahan utama dalam proses binerisasi adalah menentukan nilai ambang (*Threshold*). Nilai ini digunakan untuk mempartisi citra ke dalam dua nilai yaitu hitam (1) dan putih (0). Tujuan Metode Otsu adalah membagi histogram citra gray level ke dalam daerah yang berbeda secara otomatis tanpa membutuhkan bantuan *user* untuk memasukan nilai ambang pendekatan. Metode otsu melakukan analisis diskriminan yaitu menentukan suatu variable yang dapat membagi objek latar depan (*foreground*) dan latar belakang (*Background*) [4].

2.4 Automatic Cropping

Automatic cropping adalah proses pemotongan citra pada elemen tertentu pada area citra. Proses ini bertujuan untuk mengambil elemen yang diinginkan dari citra digital. *Automatic cropping* merupakan salah satu cara segmentasi citra dengan menggunakan histogram. Segmentasi citra adalah proses membagi suatu citra menjadi bagian-bagian yang homogen berdasarkan kriteria keserupaan tertentu antara tingkat keabuan dari piksel dengan tingkat keabuan piksel tetangganya. Citra input tulisan tangan karakter Mandarin akan dipotong untuk mendapatkan citra tulisan tangan yang telah dibuang bagian tepinya. Proses pemotongan citra input dilakukan dengan menggunakan *automatic cropping*. Tahap pertama pada *automatic cropping* adalah dengan

melakukan segmentasi baris, proses ini dilakukan dengan teknik analisis piksel dengan *vertical histogram* dimana piksel akan ditelusuri dari pojok kiri atas hingga pojok kanan bawah secara per baris. Jika baris piksel tidak mempunyai titik hitam maka informasi tersebut tidak ada [6]. Tahap kedua adalah dengan menggunakan segmentasi kolom. Proses ini dilakukan dengan cara yang sama, tetapi perhitungan piksel hitam dilakukan secara per kolom.

2.5 Normalisasi Ukuran

Normalisasi ukuran adalah suatu proses untuk menyeragamkan dimensi citra agar keakuratan dan pada baris data yang dimiliki mempunyai ukuran yang seragam. Dimensi citra yang akan digunakan adalah 32x32, karena dimensi tersebut tidak terlalu besar dan tidak terlalu kecil. Jika dimensi citra terlalu besar maka proses akan membutuhkan waktu yang lama, sebaliknya jika dimensi citra terlalu kecil akan mempengaruhi hasil dan tingkat keakuratannya.

2.6 Ekstraksi Ciri

Ekstraksi ciri merupakan bagian dari teknik pengenalan pola yang bertujuan mengambil atau mengekstraksi nilai-nilai unik dari suatu objek. Ekstraksi ciri yang digunakan dalam perancangan ini adalah *global histogram*. Histogram menunjukkan distribusi piksel berdasarkan intensitas yang dimiliki oleh tiap-tiap piksel. *Global Histogram* merupakan gabungan dari beberapa histogram, yaitu [8]:

1. Histogram Vertikal
2. Histogram Horizontal
3. Histogram Diagonal kiri
4. Histogram Diagonal kanan.

Karena pada perancangan ini digunakan skala perbandingan 32x32 maka akan didapatkan jumlah nilai hasil ekstraksi ciri (Vektor Ciri) seperti di bawah ini :

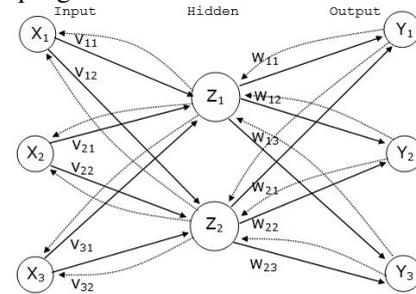
1. Jumlah Vektor Ciri Histogram Vertikal:32
2. Jumlah Vektor Ciri Histogram Horizontal:32
3. Jumlah Vektor Ciri Histogram Diagonal Kiri:63
4. Jumlah Vektor Ciri Histogram Diagonal kanan: 63

Global Histogram adalah nilai dari penjumlahan nilai vertikal, horizontal, diagonal kiri dan diagonal kanan yang berjumlah 190. Nilai dari masing-masing ekstraksi ciri seperti di atas (berupa vektor ciri) merupakan input bagi *Backpropagation Neural Network*.

2.7 Back Propagation Neural Network

Back Propagation Neural Network (BPNN) merupakan salah satu pemrosesan informasi atau data yang dibentuk dengan menirukan cara kerja otak manusia. Salah satu *Neural Network* yang banyak digunakan untuk pengenalan pola adalah *Backpropagation Neural Network*. Algoritma pembelajaran BPNN memperkecil tingkat *error* dengan cara menyesuaikan bobotnya berdasarkan perbedaan nilai output dan target yang diinginkan. Arsitektur BPNN menggunakan banyak lapisan (*multi layer*). Dengan adanya *hidden layer* dapat menyebabkan tingkat *error* yang lebih kecil [5]. Jumlah

lapisan dan jumlah neuron dalam *hidden layer* ditentukan dengan metode *try and error* untuk mendapatkan nilai *error* terkecil. Neuron di *input layer* mewakili vektor ciri dari proses ekstraksi ciri dari pola yang akan dikenali. Neuron *output layer* menunjukkan hasil dari pengenalan.



Gambar 3. Arsitektur BPNN

Setiap neuron mempunyai keadaan internal yang disebut aktivasi atau level aktivasi. Secara tipikal suatu neuron mengirimkan aktivasinya ke beberapa neuron lain sebagai sinyal. Fungsi aktivasi yang biasa digunakan untuk menentukan nilai output dari neuron di *hidden layer* dan di *output layer* pada BPNN adalah Fungsi Aktivasi Sigmoid Biner

$$y = f(x) = \frac{1}{1 + e^{-x}} \quad \dots\dots(2)$$

Dan fungsi turunannya adalah

$$y' = f'(x) = f(x)((1 - f(x))) \quad \dots\dots(3)$$

Proses pembelajaran BPNN terdiri dari 3 tahap, yaitu tahap propagasi maju, tahap propagasi balik, dan tahap penyesuaian bobot. Tahap propagasi maju digunakan untuk menghitung nilai output hasil pembelajaran. Pada tahap propagasi balik, nilai output ini akan dibandingkan dengan nilai target untuk menghitung nilai *error* pada setiap lapisan. Nilai bobot keterhubungan yang baru untuk pembelajaran berikutnya didapat pada tahap penyesuaian bobot.

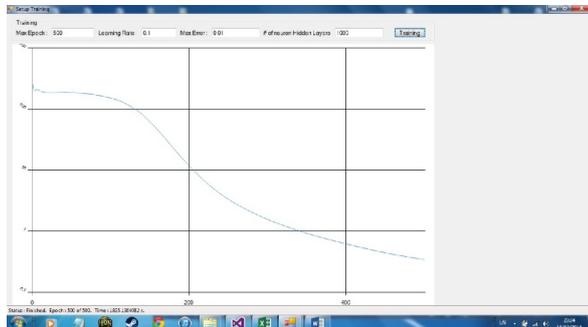
2.8 Pengujian

Pengujian program pengenalan Karakter Mandarin dengan menggunakan *Backpropagation Neural Network* (BPNN) ini terdiri dari beberapa tahap. Jumlah karakter yang diuji sebanyak 300 karakter yang masing-masing ditulis oleh : Li Ting, Shuang De, Ardy, dan Wilson. Setiap karakter dikelompokkan menurut jumlah goresan yang membentuknya seperti pada tabel 2.

Tabel 2. Pengelompokan Karakter

Jumlah Goresan	Jumlah Karakter
1-4	54
5-8	84
9-13	82
13-16	80

Tahap pertama pengujian adalah melakukan proses pelatihan dengan menggunakan data uji sebanyak 600 data latih (diambil dari tulisan Li Ting dan Shuang De). Variabel parameter BPNN yang digunakan dapat dilihat pada tampilan program menu pelatihan (gambar 4).



Gambar 4. Tampilan program menu pelatihan

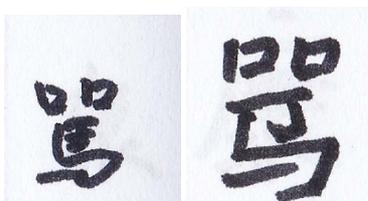
Dari hasil beberapa pelatihan tidak pernah didapatkan *error* yang maksimum, proses pelatihan selalu berhenti pada saat epoch maksimum yang memakan waktu proses sekitar 30 menit.

Tahap kedua pengujian merupakan pengenalan terhadap data latih (Li Ting dan Shuang De) dan data uji (Ardy, dan Wilson). Hasil pengujiannya dapat dilihat pada tabel 3.

Tabel 3. Hasil pengujian data latih dan data uji

Jumlah Goresan	Li Ting	Shuang De	Ardy	Wilson
1-4	98.1%	97.2%	57.4%	59.2%
5-8	96.4%	98.8%	49.9%	54.2%
9-12	97.5%	98.1%	50.6%	51.8%
13-16	98.6%	98.4%	48.7%	50%

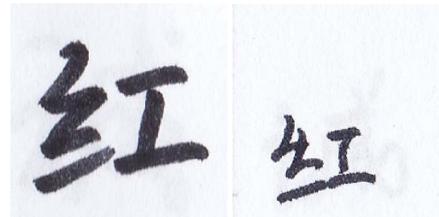
Hasil pengenalan memiliki presentase yang bervariasi, yang disebabkan oleh beberapa kesalahan dalam penulisan sehingga mengakibatkan citra karakter mandarin tersebut tidak dapat dikenali atau proses pengenalan menjadi salah. Pada proses pengenalan dengan menggunakan data latih, tingkat keakurasian dapat mencapai 98,8 % yang didapat pada karakter dengan jumlah goresan 5-8. Tingkat keakurasian pengenalan dengan menggunakan data uji hanya mencapai 59,2 % pada karakter dengan jumlah goresan 1-4. Gambar 6 menunjukkan salah satu kesalahan penulisan karakter mandarin.



Gambar 5. Kesalahan penulisan

Tebal tipis cara penulisan pada karakter mandarin yang sudah di *scan* juga menjadi masalah yang dapat mempengaruhi keakurasian dalam pengenalan, karena

ekstraksi ciri *global histogram* menghitung nilai piksel perbaris dan perkolom.



Gambar 6. Tebal Tipis Tulisan

3. Kesimpulan

Pengenalan karakter mandarin dengan metode *Backpropagation Neural Network* (BPNN) dan ekstraksi ciri *global histogram* dapat mencapai tingkat keakurasian sampai 59,2 %. Kesalahan pada pengenalan didominasi oleh adanya penebalan pada penulisan karakter mandarin sehingga menyebabkan penyebaran tinta pada kertas. Selain itu pada proses pembelajaran juga tidak didapatkan *error* minimal yang ditentukan. Untuk menjadikan pengenalan menjadi maksimal disarankan menambahkan proses *thinning* pada citra input terlebih dahulu dan mencari nilai variabel parameter BPNN yang optimum.

Daftar Pustaka

- [1] Adelia, Mengenal Bahasa Mandarin, <http://www.mandarin.web.id/>, 5 September 2014
- [2] David, Pengenalan Karakter Mandarin, <http://journal.uui.ac.id/index.php/Snati/article/view/2650/2429>
- [3] Digilib, Pengertian Smoothing, <http://digilib.its.ac.id/.../ITS-Undergraduate-5131-51041000>, 26 Oktober 2014
- [4] Erikanya Dede, Metode Otsu, <https://id.scribd.com/doc/51626189/METODE-OTSU>, 6 Oktober 2014
- [5] Fausett, Laurent. *Fundamental Of Neural Network : Architectures, Algorithms and Applications*, (Englewood Cliffs : Prentice-Hall, 1994), hal. 291
- [6] Jans Hendry, Using Profile Projection To Segmentation Character In Image (Matlab). <https://id.scribd.com/doc/67324860/Segmentasi-Karakter-Alphabet-Pada-Citra-Digital-OCR-Menggunakan-Profil-Projection11>, 11 September 2014.
- [7] Kohar, Sejarah Singkat perkembangan Bahasa cina, <http://www.tdwclub.com/showthread.php?1879-sejarah-singkat-perkembangan-bahasa-cina>, 06 september 2014
- [8] Michael Simon. *Global Histogram Sebagai Pengekstrasi Ciri Untuk Pengenalan Karakter Mandarin Program Studi Teknik Informatika Fakultas Teknologi Informasi, Universitas Tarumanagara* h. 20, 2013.
- [9] Sesepuh EOC, Sekilas tentang Bahasa Mandarin, <http://www.eocommunity.com/Sekilas-Tentang-Bahasa-Mandarin>, 7 Oktober 2014
- [10] Steve, Smoothing, <http://www.cse.psu.edu/~rcollins/CSE486/lecture04.pdf>, 12 September 2014

Biodata Penulis

Chairisni Lubis, memperoleh gelar Sarjana Fisika (Dra), Jurusan Fisika Universitas Indonesia, Depok, lulus tahun 1989. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Ilmu Komputer, Universitas Indonesia, Depok, lulus tahun 2000. Saat ini menjadi Dosen Tetap di Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

Tony, memperoleh gelar Sarjana Komputer (S.Kom), Program Studi Sistem Komputer Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta, lulus tahun 2005. Memperoleh gelar Magister Komputer (M. Kom) dari Fakultas Ilmu Komputer, Universitas Indonesia, Depok, lulus tahun 2010. Saat ini menjadi Dosen Tetap di Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

Ardy Kuncoro, memperoleh gelar Sarjana Komputer (S.Kom), Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta, lulus tahun 2015.

